A. Concepts

a. In what ways can sampling be better than a census?

There are a number of advantages of sampling over a complete census, for example:

1. Reduce cost and respondent burden.
2. Greater speed and scope, if highly trained personnel or equipment is needed.
3. Greater accuracy, because it is easier to train small crew, supervise them, and reduce data entry errors.

If a sample could be representative in the sense that characteristics of interest in the population can be estimated from the sample with a known degree of accuracy. Then it will provide clearly better data than a long census form.

b. How is Sampling different from Experimental Design?

1. In sampling, one wishes to see what the population is like without disturbing it, but in experiments, one deliberately perturbs some part of population to see the effect of the action.
2. Besides, in sampling the objective is to describe the characteristics of the population. However, in experiments, the objective is to compare the mean response to changes in levels of the population.

b. Give examples of four types of selection bias. How can samples with selection bias still be useful?

1. **Convenience sampling**: Sample was chosen for ease of selection.

   **Example**: Suppose some researchers want to find out which of two mayoral candidates is favored by voters. The researchers might ascertain the voting preferences of their own friends and acquaintances. Or they might interview shoppers at a local mall. Or they might publish two telephone numbers in the local newspaper and ask readers to call either number in order to "vote" for one of the candidates.

   **Possible effects**: The important feature of these methods is that they would systematically exclude some members of the
population (respectively, eligible voters who do not know the researchers, do not go to the shopping mall, and do not read the newspaper). Consequently, their findings could not be generalized to the population of city voters.

2. **Judgment sampling**: The investigator uses his or her own judgment to select what seems like an appropriate sample. **Example**: A TV researcher wants a quick sample of opinions about a political announcement. They stop what seems like a reasonable cross-section of people in the street to get their views. **Possible Effect**: This method is highly liable to bias and error as the researcher makes inexpert judgment and selection.

3. **Under coverage and over coverage**:  
   i. **Under coverage**: Failing to include all of the target population in the sampling frame, i.e. when some members of the population are inadequately represented in the sample. **Example**: A classic example of under coverage is the Literary Digest voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from under coverage of low-income voters, who tended to be Democrats.
   
   ii. **Over coverage**: Including population units in the sampling frame that are not in the target population. **Example**: Use of lists to develop the survey sampling frame might overlook events such as business mergers or changes in a facility’s ownership. When the survey sampling frame is created by merging several lists, consistent identifiers for each sampling unit are essential in order to discard duplicate entries.

   iii. **Possible Effect**: Both under coverage and over coverage can occur at several junctures during the survey process. For instance, in population surveys in which the sample is selected in two or more stages to obtain estimates of persons within households, coverage errors may occur at any or all stages when creating the sampling frame of primary sampling units, during field listing of housing
units or when creating a household roster of persons within a given family. Under coverage that occurs during field listing can result if members of the survey sample are excessively expensive to locate or are part of multi-unit structures, or if maps do not accurately display the sampling area.

4. **Nonresponse:**
   Nonresponse refers to the failure to obtain observations on some elements selected for the sample.

   **Example:** There are Four common sources of nonresponse:
   
   i. **not-at-Homes, Refusals, Unable to Answer and Not Found.** Not-at-homes are people who reside at the location but are away (e.g. for vacation, work) at the time when the interviewer attempted to make contact.
   
   ii. **Refusals** are persons who are contacted but refuse to answer. Refusals may result from apathy, fear of invasion of privacy or any number of reasons. Some refusals are partial, where the respondent will answer some questions but not all. Others are temporary, where the respondent will answer on the second or third contact.
   
   iii. **Unable to answer** includes persons who are too infirm or temporarily incapacitated.
   
   iv. **Not Founds** are persons who have moved and left no forwarding address, are deceased, or are not contacted due to error in the survey procedures.

   **Possible Effect:** The consequences of nonresponse can vary. As nonresponse increases, the potential for a biased sample increase. This means that the obtained responses of a probability sample may no longer be representative of the larger population. In short, response bias can reduce a probability sample to what is essentially a convenience sample and consequently, the conclusions are much weaker.

   **d. Commonly used sampling designs:**
   
   i. **Simple random sampling:** This requires that each element has an equal probability of being included in the sample and that the list of all population elements is available.
   
   ii. **Systematic sampling** is commonly used as an alternative to SRS because of its simplicity. It selects every k-th
element after a random start (between 1 and k). Its procedural tasks are simple, and the process can easily be checked, whereas it is difficult to verify SRS by examining the results.

iii. **Stratified random sampling** classifies the population elements into strata and samples separately from each stratum. It is used for several reasons: (a) The sampling variance can be reduced if strata are internally homogeneous, (b) separate estimates can be obtained for strata, (c) administration of fieldwork can be organized using strata, and (d) different sampling needs can be accommodated in separate strata.

iv. **Cluster sampling** is often a practical approach to surveys because it samples by groups (clusters) of elements rather than by individual elements. It simplifies the task of constructing sampling frames, and it reduces the survey costs.

e. Define:
   
i. **Sampling frame:**
   A sample frame is a list that includes every member of the population from which a sample is to be taken. The listing of the accessible population from which you'll draw your sample is called the sampling frame. If you were doing a phone survey and selecting names from the telephone book, the book would be your sampling frame.

ii. **Measurement error**
   Difference between the actual value of a quantity and the value obtained by a measurement. Repeating the measurement will improve (reduce) the random error (caused by the accuracy limit of the measuring instrument) but not the systemic error (caused by incorrect calibration of the measuring instrument).

iii. **Sampling error**
   In sampling contexts, the standard error is called sampling error. Sampling error gives us some idea of the precision of our statistical estimate. A low sampling error means that we had relatively less variability or range in the sampling
iv. **Non-sampling error**
A statistical error caused by human error to which a specific statistical analysis is exposed. These errors can include, but are not limited to, data entry errors, biased questions in a questionnaire, biased processing/decision making, inappropriate analysis conclusions and false information provided by respondents.

B. **Problems from the text book**

<1.13>
(1) **The target population**: The participants who attended the 2005 Joint Statistical Meeting in Minneapolis, Minnesota.

(2) **Sampling frame**: A list of the conference registrants of the 2005 Joint Statistical Meeting.

(3) **Sampling unit**: An individual.

(4) **Observation unit**: Opinions of the participants who attended the 2005 Joint Statistical Meeting.

(5) **Discussion**:

**Selection bias and Inaccuracy of responses**: the possible sources of selection bias include:

- Overcoverage: Since some of the conference registrants did not attend the 2005 Joint Statistical Meeting, it means those persons not in the target population are not screened out of the sample.

- Undercoverage: People who attended (target population) are not in the list or don’t have e-mail address or didn’t fill out e-mail address while they registered.

- Non-response: persons who are contacted but refuse to answer.

- Convenience sampling: Sample was chosen for ease of selection. It would systematically exclude some members of the population, like participants who do not read the e-mail).

Selection bias can cause Inaccuracy of responses and lead a probability sample to what is essentially a convenience sample and consequently, the conclusions are much weaker.
(1) **The target population:** Students in public schools
(2) **Sampling frame:** Five school districts in a Western state.
(3) **Sampling unit:** student
(4) **Observation:** Students’ use of the internet.
(5) **Discussion:**

**Selection bias and Inaccuracy of responses:** the possible sources of selection bias include:

- Convenience sampling: participation in this survey was voluntary. It would systematically exclude some members of the population, like students who refuse to participate.
- Judgment Sampling: This sample were selected to give a cross-section. This method might be liable to bias and error as the researcher makes inexpert judgment and selection.
- Undercoverage: Students (target population) who did not use internet in school media centers are excluded from the survey.
- Non response: Failing to obtain observations on some elements selected for the sample, like some students refused to answer.

This selection bias can cause Inaccuracy of responses and response bias can reduce a probability sample to what is essentially a convenience sample and consequently, the conclusions are much weaker.

The shape the data appears **right skewed**
b. The mean number of publications per faculty member is:

\[
\text{Mean} = \frac{1}{n} \sum_{i=1}^{50} Y_i = 1.78 ; \text{ Variance } s^2 = 7.2
\]

\[
SE = \sqrt{\frac{N-n}{N} \times \frac{S^2}{n}} = \sqrt{\frac{807-50}{807} \times \frac{7.2}{50}}=0.96*0.37947=0.367
\]

c. The \( \bar{Y} \) from (b) is 1.78 will be approximately normally distributed. Because the sample size 50 is sufficiently large.

d. The proportion of faculty members with no publications is:

\[
\hat{p} = \frac{28}{50} = 0.56 \quad n\hat{p} = 50 \times 0.56 = 28 \quad > 5
\]

95% CI is given by \( \hat{p} \pm Z_{\frac{\alpha}{2}} SE(\hat{p}) \), \( Z_{\frac{0.05}{2}} = 1.96 \)

\[
\hat{V}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = 0.93804 \times \frac{0.56 \times 0.44}{49} = 0.0047
\]

\[
SE(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = 0.0687
\]

Thus, A 95% CI for the proportion of faculty members with no publications is: 0.56 \pm 1.96 \times 0.0687 = [0.425, 0.695]

<2.14>

a. The members of the population whose occupation government are inadequately represented in this sample. This sample fails to include all of the target population in the sampling frame. So the possible cause of selection bias in this sample is under coverage.

b. The estimation of percentage of members who are female is:

\[
\hat{p} = \frac{46}{150} = 31\%
\]

95% CI is given by \( \hat{p} \pm Z_{\frac{\alpha}{2}} SE(\hat{p}) \), \( Z_{\frac{0.05}{2}} = 1.96 \)

\[
\hat{V}(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = 0.82639 \times \frac{0.31 \times 0.69}{149} = 0.0011862
\]

\[
SE(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = 0.03444
\]

Thus, A 95% CI for percentage of members who are female is: 0.31 \pm 1.96 \times 0.03444 = [24.25%, 37.75%]

c. A 95% CI for the total number of SSC members who are female is:

\[864 \times 24.25\%, 864 \times 37.75\% = [210,326]\text{(members who are female)}\]
Explain why each of the following methods will or will not result in a simple random sample.

a. It will result in a simple random sample. Because each of the 742 numbers has an equal probability of being included in the sample.
b. From the procedure, it also can result in a simple random sample. Because each of the 742 numbers still has an equal probability of being included in the sample.
c. It will result in a simple random sample. Since Tables of random numbers have the desired properties no matter how chosen from the table: by row, column, diagonal or irregularly. The 170 units have an equal probability of being included in the sample.
d. It will not result in a simple random sample. Because from the procedure, multiplying each decimal by 200 will result in that all numbers are even.
e. It will result in a simple random sample. Because each of students has an equal probability of being included in the sample.
f. It will not result in simple a simple random sample. Because the second random select will cause the unequally probabilities for students in different classes.

The probability is \( \frac{\binom{1}{1} \times \binom{99,999,999}{999}}{\left( \frac{100,000,000}{1,000} \right)} = \frac{1,000}{100,000,000} = \frac{1}{1,000,000} \) that I will be selected in the sample.

The probability is \( \frac{2,000}{\left( \frac{99,999,999}{1,000} \right)} \) that I won't be selected in any of the sample.

Suppose that there will be X samples must be selected

\[
1 - \frac{x}{\left( \frac{99,999,999}{1,000} \right)} = 50\%
\]

I selected the SRS of 150 players by SAS.

This is my SAS code

```sas
DATA S0132;
INFILE "C:\mysas\baseball.csv" DLM=',';
```
\[ \text{logsal} = \log(D); \]
\[ \text{RUN;} \]

\[ \text{PROC SURVEYSELECT data=S0132 seed=1000 method=srs n=150; out="C:\mysas\BaseballSRS";} \]
\[ \text{run;} \]

b. The distribution of \textbf{salary} appears right skewed.

![Distribution of salary](image)

The distribution of \textbf{logsal} appears approximately normal.

![Distribution of logsal](image)

c. The mean of the variable \textbf{logsal} is:

\[ \text{Mean} = \frac{1}{n} \sum_{i=1}^{150} Y_i = \frac{2071.26}{150} = 13.8084 \]

\[ \text{Variance} \quad s^2 = 1.352333 \]

\[ \text{SE(\bar{Y})} = \sqrt{\frac{N-n}{N}} \cdot \sqrt{\frac{s^2}{n}} = \sqrt{\frac{797-150}{797}} \cdot \sqrt{\frac{1.352333}{150}} \cdot 0.9 \cdot 12.247448 = 0.08555 \]
95% CI
95% CI is given by \( \bar{y} \pm \frac{Z_\alpha}{2} \text{SE}(\bar{y}) \), \( Z_\alpha = Z_{0.05} = 1.96 \)

Thus, a 95% CI for the mean of the variable \( \text{logsal} \) is:
\[
13.8084 \pm 1.96 \times 0.08555 = [13.64072, 13.97607]
\]

d. The proportion of players in the data set who are pitchers is:
\[
\frac{64}{150} = 0.426667 (\hat{p})
\]
95% CI
95% CI is given by \( \hat{p} \pm \frac{Z_\alpha}{2} \text{SE}(\hat{p}) \), \( Z_\alpha = Z_{0.05} = 1.96 \)

\[
\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} = 0.811794 \times \frac{\frac{64}{150} \times \frac{86}{150}}{149} = 0.001333
\]

\[
\text{SE}(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = 0.036507
\]

Thus, a 95% CI for the proportion of faculty members with no publications is:
\[
0.426667 \pm 1.96 \times 0.036507 = [0.3551, 0.4982]
\]
e. From the population the true mean is
\[
\text{Mean} = \frac{1}{n} \sum_{i=1}^{797} Y_i = \frac{11099.865}{797} = 13.927
\]

The proportion of players in the population data set who are pitchers is
\[
\frac{376}{797} = 0.471769
\]

c. **External Exercises**

a. i. **Target population:** The adults (18 or older) living in an apartment complex.

ii. **Elementary unit:** One of the adults (18 or older) living in an apartment complex.

iii. **Enumeration unit:** Each household

iv. **Sampling frame:** The list of households which is available at the apartment manager’s office.

v. **Variable of the interest:** exercising hours
b.

The number of all the samples of two house are \( \binom{9}{2} = 36 \)

The average exercising time per person for each sample, see the following table (the part of pink shade)

<table>
<thead>
<tr>
<th>household1</th>
<th>ad1</th>
<th>hr1</th>
<th>household2</th>
<th>ad2</th>
<th>hr2</th>
<th>total-ad</th>
<th>total-hr</th>
<th>(b)avg hr/adult/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>2.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>9</td>
<td>1.29</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0.50</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>10</td>
<td>2.00</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>1.38</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>1.17</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>15</td>
<td>2.14</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>11</td>
<td>2.20</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>2.33</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>2.00</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>1.80</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>1.75</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>1.50</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>8</td>
<td>1.33</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>1.29</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>1.25</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>1.14</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>1.00</td>
</tr>
</tbody>
</table>
c. The expected value of this estimate is $E[\bar{y}] = \frac{1}{36} \sum_{i=1}^{36} \bar{y}_i = 1.08$

Bias = $E[\bar{y}] - \bar{Y}_u = 1.08 - \frac{26}{23} = -0.05 \neq 0$

Thus this estimator is not unbiased.

d. SE = $\sqrt{\frac{\sum_{i=1}^{36} (\bar{Y}_i - \bar{Y})^2}{36-1}} = \sqrt{0.3951} = 0.6286$

SD = $\sqrt{\frac{\sum_{i=1}^{36} (\bar{Y}_i - \bar{Y}_u)^2}{36-1}} = \sqrt{0.3977} = 0.6306$

SD − SE = 0.002